

Název práce: Sufixové grafy a bezeztrátová komprese dat

Autor: Martin Senft

Katedra: Katedra software a výuky informatiky

Vedoucí doktorské práce: doc. RNDr. Tomáš Dvořák, CSc., Katedra software a výuky informatiky

Abstrakt: Sufixový strom a příbuzné datové struktury umožňují asymptoticky optimálně řešit řadu úloh o řetězcích a jejich vlastnostech lze též využít k implementaci metod bezztrátové komprese dat. Cílem práce je prozkoumat možnosti opačného přístupu, tedy využití vlastností sufixových grafů k návrhu kompresních algoritmů. Práce popisuje univerzální konstrukční algoritmus pro sufixový trie, sufixový strom, DAWG a CDAWG, doprovázený analýzou simulace implicitních sufixových hran, která přináší dvě praktické alternativy k tradičnímu řešení. Protože kompresní metody vyžadují udržování textu v posuvném okně, je třeba rozebrat chování sufixových grafů v této situaci. V práci je ověřeno, že pouze sufixový strom je schopen udržovat posuvné okno v amortizovaně konstantním čase, zatímco CDAWG (podobně jako DAWG) vyžaduje čas úměrný délce okna, což řeší hypotézu Inenagy a kol. Na tomto základě je popsána třída kompresních algoritmů, založených pouze na popisu konstrukce sufixového grafu nad komprimovaným textem. Zatímco některé z algoritmů odpovídají klasickým slovníkovým či kontextovým metodám bezztrátové komprese, jiné jsou zcela originální. Výklad je doplněn pilotní implementací navržených algoritmů a experimentálním vyhodnocením na standardních datových korpusech.

Klíčová slova: sufixový strom, CDAWG, konstrukce, posuvné okno, komprese dat